P2P 系统中基于 Web 的索引信息扩展算法的研究

尤佳莉[1,2],王劲林[1]
YOU Jia-li[1,2],WANG Jin-lin[1]

1.中国科学院 声学研究所,北京 100080
2.中国科学院 研究生院,北京 100080
1.Institute of Acoustics,Chinese Academy of Sciences,Beijing 100080,China
2.Graduate School of Chinese Academy of Sciences,Beijing 100080,China
E-mail:youjiali@mails.gucas.ac.cn

YOU Jia-li,WANG Jin-lin.Algorithm of Web based index expansion in P2P systems.Computer Engineering and Applications, 2008,44(18):6-8.

Abstract: Based on the characters of Web and Peer-to-Peer(P2P) systems,an algorithm of auto-generating index of media files for P2P systems from Web is proposed.This method can efficiently solve the problem about lack of descriptors of media files, which causes the low recall and precision in key words searching.Besides,to satisfy the increasing information of Web and the changing popularities of people,an auto-updating strategy for peers is applied.From the experiments,it shows that this method can expand the describing information for media files and significantly improve the searching capacity of P2P systems.
Key words: Peer-to-Peer;Web;index;search

摘 要:基于环球网(Web)的特点和用户在点对点(P2P)系统中搜索的习惯,提出了一个在 P2P 系统中对媒体文件自动生成索引的方法。该方法有效地解决了媒体文件描述符不足所带来的查询精度低的问题。同时,提出了一个在 P2P 系统中节点信息的更新策略。实验表明,描述符扩展后,媒体文件查询结果的准确率得到了显著的提高。
关键词:点对点;环球网;索引;查询
DOI:10.3778/j.issn.1002-8331.2008.18.002 **文章编号:**1002-8331(2008)18-0006-03 **文献标识码:**A **中图分类号:**TP301

1 前言

近年来,对等网络技术的发展,使得网络信息传输和共享更加灵活和方便。在当前对等网络的应用中,音视频等媒体文件的检索和传输占据了很重要的位置[1],而对应的视频流媒体技术也得到了快速的发展[2,3]。随着影音等内容的不断增多,怎样快速、精确地找到用户所感兴趣的内容,以及给用户一些有效的建议,是一个很有潜力并颇具挑战的研究方向。在当前的大多数 P2P 系统中,直接而有效的方法是对关键词建立倒排索引[4-9]。现阶段,对于电影来说,在终端节点上对其音视频进行基于内容的信息提取是不现实的,一般可以得到的描述信息仅仅有文件名字,通常是片名。而对于一个影片,抽象化的名字难以表达电影的内容信息。在查询时,如果用户不能精确地输入片名,则很难找到想要的结果。因此,描述信息的不足,是媒体文件查询性能低的一个重要因素[10]。

Web 可以看作是一个巨大的、包含各种信息的数据库[11]。由于交互自由和信息开放,对于电影等媒体文件,很多电影公司和用户会在 Web 上发布一些描述性信息, 如果在对 P2P 系统中的文件建立索引时能正确地挖掘这些信息,则可能会有效地扩充媒体文件的描述符,提高查询性能。因此,本文提出了一个 P2P 系统中基于 Web 的索引信息自动扩展的方法来扩充媒体文件的描述符,同时针对结构化 P2P 系统(Distributed Hash Table,DHT)的特点,提出了一个索引信息自动更新的策略,从而提高查询的正确性。在本文的实验中,根据用户查询习惯,自动生成查询问题(Query)进行测试。实验结果表明,索引信息扩展后,系统的查询性能得到了显著提高。

2 相关工作

在结构化 P2P 系统中,基于关键词的精确查询得到广泛应用,其基本思想是:首先对文件中的关键词进行提取,计算其在当前文件中的统计信息,并对关键词建立倒排索引表,并用分布式哈希的方法将该表平均分布在整个网络中。倒排索引表的结构如图 1 所示。

对于这个方法,许多人进行了深入地研究[4-9]。 文献[4]中介绍了一个文本文件的全文检索系统,在这里,文本中的词被

基金项目:中国下一代互联网示范工程(Supported by China Next Generation Internet Foundations(CNGI) No.CNGI-04-15-2A);国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2005AA1032)。
作者简介:尤佳莉(1982-),女,博士生,主要研究领域宽带多媒体通信;王劲林(1964-),男,博士生导师,主任研究员,主要研究领域宽带多媒体通信。
收稿日期:2008-02-18 **修回日期:**2008-04-14

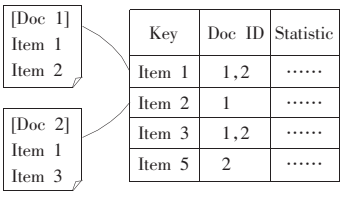

图 1　关键词倒排索引表实例

去噪提出,并建立倒排索引表。当用户进行查询时,对查询问题中的关键词逐个进行查询,将所有关键词返回结果中的相同部分作为结果,展现给用户。同样,文献[5]也是一个相似的系统,其中,通过计算常常被查询到的词在文本中的权重,来表示当前词对文档的重要程度,并对这些词进行选择并建立索引。因此每个文档都看成一组索引词的集合,这些索引词在语义上代表了文档的内容。当用户进行关键词检索时,对每个关键词进行单独检索,返回的是一个文档 ID 集合,其中包括关键词和每个文档的权重值,根据式(1)来计算查询的所有关键词和每个文档的相关性,对其排序得到检索结果。其中,d_i 表示任意文档,Q 表示 Query 的关键词集合,$s_{Q,j}$ 是第 j 个关键词在 Query 中的权重,$s_{i,j}$ 是第 j 个关键词在文档 d_i 中的权重。

$$sim(Q,d_i)=\frac{\sum_{i=1}^{n} s_{Q,j}\times s_{i,j}}{\sqrt{\text{number of terms in } d_i}} \tag{1}$$

为了减少每个节点对倒排索引表的存储消耗以及索引表传输时的通信量,文献[6]提出了一个在查询时只返回跟查询相关性高的节点,并在节点中进行详细查询的方法。由于电影等媒体文件的相关描述信息非常少,而不同终端上的同一个文件副本可能存在不同的描述信息,因此,在文献[10]中,作者提出了一个自查询的方法,逐渐发掘在一个 P2P 网络中与一个文件相关的描述信息,从而对其描述符进行扩充。

在媒体文件的描述信息中,往往只有文件名或者作者等简单的内容,因此很难从频率上进行统计,计算关键词与文件的相关性。在实际应用中,人们常常会忘记影片的片名,而只记得影片中的一些关键信息,如演员、主角、背景以及剧情描述信息等等,但这些内容并没有包含在文件的描述符中。这个问题一方面带来了检索错误,另一方面使得当用户对影片名字不能精确确定时,难以通过其他信息来找到相同或者相似内容的电影。针对以上问题,本文提出了通过对 Web 中相关信息的提取来扩展文件描述符的方法,从而提高检索性能。

3　基于 Web 的索引关键词自动生成算法

对于一些媒体文件(以电影为例),很多组织、机构或者个人会发布一些相关信息,如导演、演员、内容介绍、评价、字幕等等,而这些信息在网页中常常会和一些固定的词共同出现,比如"电影"、"电视剧"。如果能够正确、合理地抽取以上信息,则可以有效的扩充电影的描述符。当前,集中式的搜索引擎得到了很大的发展(如 Google、Baidu、Yahoo 等),这些搜索引擎都提供开放的接口供用户使用,通过动态更新搜索引擎的数据库,Web 中内容的变化可以被搜索引擎及时发现,因而搜索结果也符合整个网络的变化。在当前的 P2P 系统中,大多数电影的描述符仅仅是文件名(片名),而缺乏内容相关信息。如果用集中式搜索引擎对这些片名和固定词(如"电影")进行联合搜索,并对返回的前 N 个(Top-N)结果中链接的内容进行分析和提取,则很有可能会得到一些反映影片内容或特点的描述信息。

然而,在 Top-N 结果中,存在两个问题:第一,存在很多噪声。每个网页不仅仅包含对影片的描述信息,同样还包括广告、其他影片推荐和用户发表的不相关评论等无效信息;第二,在搜索到的影片内容中,大多用比较简短精练的语言进行描述,很多词只出现一两次,即使每个词和影片的相关性不相同,但也很难从频率上进行区分。因此,怎样除去噪声并提取和影片相关性强的词作为描述符,是一个关键问题。

互信息常常被用来描述两个词之间的共现关系,从而表示其语义相似性。如果存在一个巨大的、几乎包含所有文本的集合,词 w_1 和 w_2 的互信息 $MI_1(w_1,w_2)$ 可以如式(2)所示,其中 $C(w)$ 表示 w 在集合中出现的频率,N 表示集合中所有词的频率总和。

$$MI_1(w_1,w_2)=\text{lb}\frac{P(w_1,w_2)}{P(w_1)P(w_2)}=\text{lb}\frac{C(w_1,w_2)/N}{C(w_1)/N*C(w_2)/N} \tag{2}$$

同样,一个影片片名和一个词的互信息,可以用来衡量影片和这个词的相关性。在实际中,虽然不存在这样一个包罗万象的集合,但 Web 提供新的机会,怎样通过搜索引擎从 Web 上找到有用的内容,则是要解决的问题。

本文提出了一个基于 Web 信息的描述符提取算法,该算法由两部分组成:提取候选关键词;计算候选关键词与文件的相关性并得到文件的描述信息。下面将对以上两步进行详细介绍。

(1)提取候选关键词

由于电影或其它音视频文件的介绍信息常常会和一些类别词共同出现,如"电影"、"电视"、"音乐"等,因此,如果对片名和文件类别词用搜索引擎进行联合查询,则会增加返回相关文本内容的概率,因此,提取候选关键词的过程如下:

①对"片名"+"电影"进行联合查询;

②对 Top-N 结果中的内容进行提取,分词,去除停用词,并根据式(2)计算所有词与片名的 MI 值;

③根据 MI 值对所有词进行排序,选择 MI 最高的 M 个词作为候选关键词。

(2)计算候选关键词与文件的相关性并得到文件的描述信息

由于候选关键词仅从 Top-N 的结果中抽取,所计算的 MI 值只是一个粗略表示。因此,用 Web 信息对 MI 值进行重新计算和提纯。由于当前的搜索引擎并不能返回关键词 w_i 的频率,而是存在该词的网页的数量(即 Page Count)$C_p(w_i)$,因此式(2)可以改写为式(3)。

$$MI_2(f,w)=\text{lb}\frac{P(f,w)}{P(f)P(w)}=\text{lb}\frac{N_pC_p(f,w)}{N_pC_p(f)N_pC_p(w)} \tag{3}$$

其中,$C_p(f,w)$ 表示片名 f 和词 w 共同出现的 Page Count;N_p 表示用搜索引擎可以搜到的所有中文网页的数量。但是,现在常用的搜索引擎并不提供这个数值,如 Google。因此,需要对 N_p 进行预测。在中文中,词是无穷无尽的,几乎没有一个集合可以包含所有的词。但是,有一些功能词,如"的"、"在"、"有"等等,几乎出现在每一个中文网页上。如果对这些词的并集进行搜索[12],得到的值可以用来估计搜索引擎可以搜到的所有中文网页的数量。因此

$$MI(f,w)=MI_2(f,w) \tag{4}$$

这里,选择 Top-20 的结果进行分析。

4　基于自动更新索引的 DHT 网络

在当前的很多结构化 P2P 网络中,每个节点维护一个索引词倒排表[4-7],其中包括索引词、所对应的文件 ID 以及文件与索引词的统计信息。本文主要针对媒体文件的点播或者下载,因此,需要统计的信息包括 3 部分:片名和索引词的互信息值、影片被选中点播或者下载的次数、影片在 Web 上的频率信息(通过搜索引擎得到 Page Count)。这里提出一个策略帮助系统对统计信息进行自动更新。其中,每个节点都扩充了一个邻居表,其中存储了 M 个邻居节点,作为自动更新时的辅助计算节点。

当有新节点加入网络时,为了使网络尽快得知新节点信息,新节点仅仅对本地影片的片名进行分词,去除停用词,并计算本地影片名中的词和片名的互信息,以及在 Internet 上搜索到的该影片的 Page Count,以此作为影片的索引信息加入网络。之后,分析上述得到的 Top-N 的搜索结果,得到一个候选词集合。为了快速扩充更多的索引信息,在该节点加入网络后,它将影片的候选词集合随机分发给邻居表中的节点,由本地节点和邻居节点共同计算候选词和片名互信息。每个候选词计算完成后由计算节点将候选词、文件 ID、互信息值和 Page Count 返回给新节点,由新节点将筛选后的词提交给 DHT 网络。

当节点空闲时,则自动根据索引表找到超过一定更新时间并且被点播或者下载频率小于某个预先设定阈值的所有影片名,并对其进行搜索,重新从 Web 上抽取描述信息。计算完成后,对描述符进行依次查询并进行更新。如果新生成的索引项在 DHT 网络中不存在,则将此项加入。

当进行多关键词查询($KW=\{k_1,k_2,\cdots,k_n\}$)时,系统返回每个关键词 k_i 的查询结果,本地节点对返回的所有信息进行计算和排序。计算的方法如式(5)所示:

$$Score(f,KW)=\sum_{i=1}^{n}MI(f,k_i)+\alpha Pop\_p2p(f)+\beta Pop\_Global(f)$$

$$Pop\_Global(f)=\frac{PageCount_f}{\sum_{f_i}PageCount_{f_i}} \tag{5}$$

$$Pop\_p2p(f)=\frac{HitCount_f}{\sum_{f_i}HitCount_{f_i}}$$

其中:f 表示对应的影片,$Pop\_p2p(f)$ 表示 f 在当前 P2P 网络中的流行度;$Pop\_Global(f)$ 表示 f 在 Internet 中的流行度,f_i 表示返回的所有片名。α 和 β 是权重,分别用于调整每一项对结果的影响。由于很难在短时间的实验中获得 $Pop\_p2p(f)$ 的值,这里选取 α=0。

5　实验结果

5.1　实验数据的准备和分析

实验中,加入了近 20 000 部影片。为了测试基于内容关键词查询的正确率,根据影片类型,按比例选出 100 个影片作为测试集,并人工对内容标记关键词。为避免片面性,在这里让两个人同时进行标记,取并集作为关键词。从标记的结果中发现:

(1)题目中的词很少在被选出的关键词中;

(2)很多人物、地理、时间和影片特点等信息被选出,而这些是片名所不能包含的。

上述分析说明,题目并不能很好地表示影片内容;当不知道精确的片名时,用户可能会尽量多地输入相关词进行查询,这种情况下,在仅仅对片名等少量信息进行索引的系统中,很难找到合适的结果。

在实验中,根据人工选出的关键词随机生成共 6 707 个 Query。在实际应用中,Query 中关键词的数目分布并不均匀[1]。生成的 Query 中关键词数目分布如图 2 所示。

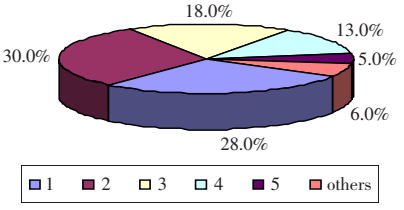

图 2　Query 中关键词数目的分布(引自文献[1])

5.2　检索性能评估和分析

MRR 是一个在查询中常用的评估准则[13],定义如式(6)所示:

$$MRR=\frac{\sum_{i=1}^{N}\frac{1}{rank_i}}{N} \tag{6}$$

其中,N 是 Query 的数目,$rank_i$ 表示第 i 个 Query 的结果中正确结果所对应的排名 rank。由于 MRR 是一个适用于衡量单一、特定查询结果的准则,因此,本文对描述符扩展前后的 MRR 值进行比较,来评估系统的查询性能。其结果如图 3 所示。

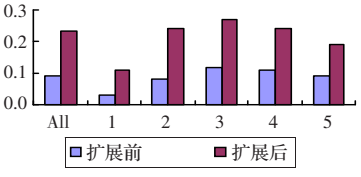

图 3　扩展前后 MRR 值比较

(All 表示所有 Query 问题的查询结果;1 表示仅含有 1 个关键词的 Query 的查询结果;2 表示仅含有 2 个关键词的 Query 的查询结果;其余类似。)

同时,描述符扩展前后的索引条目对比如表 1 所示。

表 1　描述符扩展前后索引文件的比较

	索引词数目	每个索引词对应的平均文件数
扩展前	17 168	5
扩展后	59 549	71

从表 1 中可以看到,关键词扩展后,索引词的条目得到了极大的扩充,同时也使得一个词与更多的文件相关。尽管更多索引条目会造成搜索一个词语出现大量候选答案的情况,但从图 4 可知,通过本文的算法,仍可以很好地对候选答案进行排序和区分,使 MRR 值得到了显著的提高,改善了内容关键词的查询正确率。同时,也可以看出,当 Query 中关键词个数只为 1 时,很多相关内容不能正确找到,主要原因是查询条件过于简单,查询结果缺乏有效的区分性。而当 Query 中关键词个数大于 3 时,性能也有所下降,原因是有些查询关键词是人名、地名等专有名词,而这些没有被加入到索引文件中。

虽然描述符扩展后,系统的查询性能有所提高,但是如果

(下转 24 页)